

Characterizing Multivariate Information Flows

Shohei Hidaka*

*Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

Submitted to Proceedings of the National Academy of Sciences of the United States of America

One of the crucial steps in scientific studies is to specify dependent relationships among factors in a system of interest. Given little knowledge of a system, can we characterize the underlying dependent relationships through observation of its temporal behaviors? In multivariate systems, there are potentially many possible dependent structures confusable with each other, and it may cause false detection of illusory dependency between unrelated factors. The present study proposes a new information-theoretic measure with consideration to such potential multivariate relationships. The proposed measure, called multivariate transfer entropy, is an extension of transfer entropy, a measure of temporal predictability. In the simulations and empirical studies, we demonstrated that the proposed measure characterized the latent dependent relationships in unknown dynamical systems more accurately than its alternative measure.

Time Series Analysis | Information Theory | System characterization

One of crucial steps in scientific studies is the characterization of a system of interest - specification of dependent relationships among factors or subcomponents in the system [1]. The characterization is often the early stage of analysis before proceeding to a more specific description of the systems, and it requires little or no prior knowledge of the underlying mechanism of the system. In the present study, we propose a measure which may serve as such an early-stage characterization of dependency in a multivariate system with little knowledge through observation of its temporal behaviors.

A basic problem concerned in the present study is how we can quantify and detect dependency between each pair of variables in a multivariate system. In a multivariate system, a variable X generated by a stochastic or a deterministic process is said to be conditionally dependent of variable Y given Z if its future state of X is partially determined by another variable Y given Z . In particular, we consider temporal behaviors of a system in which we measure dependency from a set of N variables at time t to the set of N variables at time $t + 1$. In this formulation, the temporal dependency of a system is characterized with N^2 conditional dependency between X_t and Y_{t+1} given Z_t .

In a linear system, which can be decomposed into separable subcomponents without interaction among them, characterization of temporal dependency is straightforward. For stationary linear processes, auto- and cross-correlation sufficiently characterizes a set of its linear properties. In contrast, a nonlinear system with or without a stochastic component, which is not decomposable to subcomponents, needs to be characterized with consideration to its interaction among subcomponents. A set of information theoretic measures has been proposed as a nonlinear counterpart of auto- and cross-correlation for a nonlinear system [2]. One of such information-theoretic measures, transfer entropy, has found a wide range of applications and has successfully characterized many empirical systems [3]. In the present study, we propose an extended information-theoretic measure, called *multivariate transfer entropy* (MTE), for characterization of multivariate dependency. The proposed measure is a natural generalization of transfer entropy, and it concerns potential confounding relationships among three or more variables in multivariate temporal dynamics. Thus we illustrate the extended measure after a brief overview of the related development of the information theory.

Information Theory. Ever since its establishment by Shannon [4], information theory has played a crucial role in mathematical modeling of communication. His original formulation concerns a unidirectional

information transmission through a noisy channel. In the formulation, a message X generated by an information source is sent to a receiver. The receiver receives it as a message Y through a stochastic channel in which the original message X may be changed to Y by a certain chance. This unidirectional information transmission is described with the mathematical concept *entropy* and *mutual information* of probabilistic distribution of the sent and received messages X and Y . Entropy quantifies the amount of stochastic uncertainty of an information source by the length of codes encoding the message generated by the information source. Mutual information quantifies the relative difference between uncertainty in the two ways of coding, the random variable Y alone and the variable Y with additional knowledge of another variable X . The mutual information is maximized when $X = Y$ (noiseless channel), and it is zero at the minimum when random variables X and Y are independent. Thus, mutual information characterizes the properties of the noisy channel between X and Y . Mutual information gives a mathematical ground for communication theory concerning the design of an optimal information channel given constraints. Introducing the concepts of entropy, mutual information and its variants, information theory covers and connects a wide range of fields and problems such as nonlinear dynamical system, thermodynamics, electrical engineering, probability theory, statistics, mathematics, economics, computer science, and philosophy of science [5].

Despite the mathematical elegance and many successful applications, in 1973 Shannon had pointed out the theoretical limitation of the unidirectional information transmission, and gave a prospect for an extension to information theory *with feedback* [6]. Indeed in the very same year, Marko [7] proposed the extended bidirectional information network as suggested by Shannon. In his bidirectional information network, two information sources send and receive information, and its efficiency of communication is characterized with loss of information from the bidirectional information transfer. Although both Shannon and Marko highlighted the importance of bidirectional information, it has not been well-recognized in the fields until 1990s [6]. More recently, Kantz and Schreiber [2, 3] have reintroduced a *directed* measure of statistical dependency, called transfer entropy, which is a subset of Marko's bidirectional information theory. One of the major advantages of the transfer entropy or bidirectional information transmission over Shannon's unidirectional measure is that it enables us to distinguish which factor leads or follows another separately from the other direction. After the reintroduction of transfer entropy, it has found applications in various research fields. Such successful applications include not only engineering fields relevant to information theory but also various kinds of scientific research: detecting directed dependency in cellular automata [8], machine learning [9], chemical process [10], health monitoring [11], analysis of

Reserved for Publication Footnotes

brain activity [12, 13], stock markets [14, 15], ecological monitoring programs [16], music analysis [17], and human-human/robot communication [18, 19, 20].

Limitation of the transfer entropy. Despite its successful applications, a potential limitation of transfer entropy has not been well recognized. A naive application of transfer entropy to three or more variables may cause an inaccurate characterization of a system. In the present study, we demonstrated this limitation, and we propose multivariate transfer entropy (MTE), a further extension of Marko's bidirectional information theory, as a solution. The MTE is concerned with the potential confounding relationship among three or more variables which transfer entropy does not count. Thus it extends its usability to more general situations with arbitrary topological structure of dependency among N variables. In this regard, the MTE is a nonlinear analogue of partial correlation which cancels linear confounding effects of other than correlation between the focal paired variables. In order to explicitly distinguish from the MTE, hereafter we refer to the original one as *pairwise* transfer entropy (PTE), a special case for a bivariate system. In the following section, we give a formal description of mutual information, transfer entropy and their relationship.

Mutual information and transfer entropy

Consider a unidirectional information transmission through a noisy channel in which a message X generated by an information source is sent to a receiver. Let $p(x_i)$ be the probabilistic distribution of a message $X = x_i \in \mathcal{M} = \{1, 2, \dots, M\}$, where \mathcal{M} is the set of M alphabets, and a series of messages is drawn from the probabilistic distribution by the information source. Then entropy, $-\sum_i p(x_i) \log p(x_i)$, gives the asymptotic minimum average code length by assigning the code with length $-\log p(x_i)$ to a infinitely long series of messages $\{X_1, X_2, \dots\}$.

Then suppose that we assign a code set Q of length $-\log q(x_i)$, instead of the minimum code set P of length $-\log p(x_i)$, to the message x_i with probability $p(x_i)$. Its average relative difference between the code length of Q and P called relative entropy or Kullback-Liebler divergence, $D(P||Q) = -\sum_i p(x_i)(\log q(x_i) - \log p(x_i))$, can be treated as the amount of coding error or the difference in stochastic uncertainty of $q(X)$ relative to $p(X)$. In the unidirectional information transmission described above, the entropy $H(Y)$ quantifies the amount of stochastic uncertainty of the received message Y . Likewise, the conditional entropy, $H(Y|X) = -\sum_{i,j} p(x_i, y_j) \log p(y_j|x_i)$, quantifies that of Y on knowledge of X . Then mutual information $I(X; Y) = H(H) - H(Y|X)$ is defined as difference of the entropy of Y relative to conditional entropy $H(Y|X)$ or symmetrically that of the entropy X relative to conditional entropy $H(X|Y)$. Mutual information can be interpreted as the amount of information gain by obtaining shorter code length $H(Y|X)$, Y with additional knowledge of X relative to the code length $H(Y)$ of Y alone.

Transfer entropy and bidirectional information network. Marko [7] has given a reinterpretation of the Shannon's unidirectional transmission as a network of information flows, and showed its bidirectional extension. Figure 1a depicts the unidirectional information transmission interpreted as a network flows¹. In the network, the mutual information $I(X^T; Y^T)$, as some amount of entropy of the information source X^T , is flown into the Y^T . The entropy of received message Y^T is the sum of the in-coming flow $I(X^T; Y^T)$ and the uncertainty of Y^T alone without X^T , the conditional entropy $H(Y^T|X^T)$.

This unidirectional information network is a special case of the bidirectional network (Figure 1b). We follow Marko's terminological conventions except for the term entropy rate and (pairwise) transfer entropy, which have become more standard after [5, 3]. Unlike the

one originally proposed by [7], the following formulation needs to assume neither stationary nor Markovity of time series in theory².

Suppose we have two series of random variables $\bar{X}^T = \{X^1, X^2, \dots, X^T\}$ and $\bar{Y}^T = \{Y^1, Y^2, \dots, Y^T\}$ over discrete time $t = 1, 2, \dots, T$ where the top bar \bar{X}^T means the set of random variables with superscript specifying time indices from time 1 to time T . As in the unidirectional communication, we start with a measure of uncertainty in a single variable. Entropy rate H_X^T is the sum of conditional entropies of X^t given its past states \bar{X}^{t-1} ($t = 1, 2, \dots, T-1$) [5].

$$H_X^T \equiv \sum_{t=1}^T H(X^t|\bar{X}^{t-1}) \quad [1]$$

where $H(X^t|\emptyset) = H(X^t)$ and $X^t = \emptyset$ for $t < 1$. The entropy rate H_X^T/T is the average increase at each step in the entropy of variable X by normalizing with length of time series T . Similarly, the sum of uncertainties of a random variable X at time t conditioned on knowledge of the past states of $\{\bar{X}^{t-1}, \bar{Y}^{t-1}\}$ for $t = 1, 2, \dots, T$ is called *free entropy* F_X^T . Formally, we define as follows.

$$F_X^T = \sum_{t=1}^T H(X^t|\bar{X}^{t-1}, \bar{Y}^{t-1}) \quad [2]$$

Similarly, we write $F_Y^T = \sum_{t=1}^T H(Y^t|\bar{X}^{t-1}, \bar{Y}^{t-1})$. The pairwise transfer entropy from Y to X at time T is defined as the sum of reductions in uncertainty of X^t conditional on knowledge of the past states of two variables $\{\bar{X}^{t-1}, \bar{Y}^{t-1}\}$ for $t = 1, 2, \dots, T$.

$$T_{Y \rightarrow X}^T = H_X^T - F_X^T = \sum_{t=1}^T I(X^t; \bar{Y}^{t-1}|\bar{X}^{t-1}) \quad [3]$$

where $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$ is conditional mutual information between X and Y given Z , and $I(X; Y|\emptyset) = I(X; Y)$. Similarly, $T_{X \rightarrow Y}^T = H_Y^T - F_Y^T$, and $T_{X \rightarrow Y}^T \neq T_{Y \rightarrow X}^T$ in general. Transfer entropy can be interpreted as directed "information transmission" from Y to X , since $H(X^t|\bar{X}^{t-1}) \geq H(X^t|\bar{X}^{t-1}, \bar{Y}^{t-1})$ and $H(Y^t|\bar{X}^{t-1}) \geq H(Y^t|\bar{X}^{t-1}, \bar{Y}^{t-1})$ if and only if the series of variable $\{\bar{X}^t\}$ is independent of the past states of another variable $\{\bar{Y}^{t-1}\}$ for $t = 1, 2, \dots, T$.

Network properties of transfer entropy. Marko [7] has pointed out that the relationship between entropy rate, free entropy, and transfer entropy can be viewed as a bidirectional information network (Figure 1b). The bidirectional network has two variables X and Y which send and receive messages between them. Each directed edge in the network reflects an information flow with non-negative value of corresponding entropy rate (solid line), free entropy (solid line) or transfer entropy (broken line). In each node, the total amount of in-coming information flows is identical to the total amount of out-going ones (Kirchhoff's current law). The entropy rate H_X^T is the sum of free entropy F_X^T (new information at T) and transfer entropy $T_{Y \rightarrow X}^T$ (information from another variable the past states up to $T-1$). A certain part $T_{X \rightarrow Y}^T$ of entropy rate H_X^T is transferred to Y , and the rest, called *residual entropy* $R_X^T = H_X^T - T_{X \rightarrow Y}^T$, is flown out of the network. Similarly, information is transferred from the variable Y to X . At each node and edge in the bidirectional network for the two variable X and Y , two properties, non-negativity of information flows and Kirchhoff's current law, are held. We refer these to two properties as network constraints. In order for all

¹ In the network, the superscript T is a set of time indices, and $X^T = X^1, X^2, \dots, X^T$ is a group of variables in which we obtain the Shannon's entropy by identifying $X = X^T$.

² However, estimation of transfer entropy often requires these properties due to finite sample size of dataset in practice. Marko [7] considered the limit of infinite long time series, which we can obtain by making $T \rightarrow \infty$ in the present formulation of time series of a finite length. Specifically, in this paper, we work simply with the quantities H_X^T , since its limit is not essential in the arguments presented here.

the information flows to be non-negative, it needs to satisfy the following inequality: $R_X^T = H_X^T - T_{X \rightarrow Y}^T \geq 0$ and symmetrically $R_Y^T = H_Y^T - T_{Y \rightarrow X}^T \geq 0$. In [7], an even better inequality as follows has been suggested without a proof.

$$\min(H_X^t, H_Y^t) \geq T_{X \rightarrow Y}^t + T_{Y \rightarrow X}^t \quad [4]$$

We will prove a more general version of this inequality for N -variable system ($N > 2$) in this study.

Transfer entropy as decomposition of mutual information. Another property of the PTE is as a partial factor decomposing mutual information with certain residuals³.

$$I(\bar{X}^T; \bar{Y}^T) = T_{X \rightarrow Y} + T_{Y \rightarrow X} + R_{X,Y} \quad [5]$$

where $R_{X,Y} = \sum_{i=1}^N I(x_i; y_i | \bar{X}_{i-1}, \bar{Y}_{i-1}) \geq 0$ is non-negative due to non-negativity of conditional mutual information. Equation 5 explicitly states PTE is an extension of mutual information which is a special case without feedback $T_{Y \rightarrow X} + R_{X,Y} = 0$ or $T_{X \rightarrow Y} + R_{X,Y} = 0$.

Multivariate bidirectional information network

Here we outline an extension of the pairwise transfer entropy to more general cases with three or more information sources. Technically, there are many potentially possible multivariate extensions of PTE. However, the proposed extension is justified not just by applicability to multivariate dependency but also by holding the two properties as well as PTE. Analogous to PTE decomposing of mutual information, MTE decomposes total correlation, a multivariate extension of mutual information [21] (also called multivariate constraint [22] or multiinformation [23]). Also we can view MTE as a part of bidirectional information network among N variables with non-negative flow holding Kirchhoff's current law. In the following sections, we formulate a multivariate information network and overview its theoretical properties. See also Supplemental Information for the more detailed description and the mathematical proofs of the theorems.

Formulation of multivariate information network. In a generalized information network with N variables, each variable is associated with the two nodes – in-coming and out-going node (Figure 1d). The in-coming and out-going node of variable i respectively receives and sends information from all the variables but the variable i . Information flow in between the in-coming and out-going node of variable i is the entropy rate of variable and free entropy of variable i . At the out-going node, there is some amount of information lost without being transferred to the other variables which is called residual entropy. A special case of the information network for a three variable system is shown in Figure 1c.

Let \bar{X}_N^T be a set of $N \times T$ random variables indexed with the index set $\mathcal{N} = \{1, 2, \dots, N\}$ and the index set for time $\mathcal{T} = \{1, 2, \dots, T\}$. Then let us denote $\bar{X}_N^T = \{\bar{X}_1^T, \bar{X}_2^T, \dots, \bar{X}_N^T\} = \{\bar{X}_N^1, \bar{X}_N^2, \dots, \bar{X}_N^T\}$ where $\bar{X}_i^T = \{X_i^1, X_i^2, \dots, X_i^T\}$ is time-cumulative subset of X_i for time index \mathcal{T} , and $\bar{X}_N^t = \{X_1^t, X_2^t, \dots, X_N^t\}$ is a set for the variable set \mathcal{N} given t . Given the set of random variables \bar{X}_N^T , entropy rate, free entropy, multivariate transfer entropy and residual entropy are defined as follows. The cumulative sum of entropy rates of variable i at time T is defined as

$$H_i^T \equiv \sum_{t=1}^T H(X_i^t | \bar{X}_i^{\bar{t} \setminus t}) \quad [6]$$

where $\bar{t} = \{1, 2, \dots, t\}$ is the cumulative set of time indices and $\bar{t} \setminus t = \{1, 2, \dots, t-1\}$ means set subtraction of index t from \bar{t} with the set subtract operator “ \setminus ”. This is identical to the entropy rate

defined in a bivariate system [7, 3]. Free entropy of variable i at time T in the N -variable system, which is uncertainty of x_i given all the past states of N variables $\bar{X}_N^{T \setminus T}$, is defined as follows.

$$F_i^T \equiv \sum_{t=1}^T H(X_i^t | \bar{X}_N^{T \setminus t}) \quad [7]$$

Multivariate transfer entropy from variable j to i given the set of the other variables $\mathcal{N} \setminus \{i, j\}$ in the N -variable system \bar{X}_N^T is defined as follows.

$$T_{j \rightarrow i | \mathcal{N} \setminus \{i, j\}}^T \equiv \sum_{t=1}^T I(X_i^t; \bar{X}_j^{\bar{t} \setminus t} | \bar{X}_{\mathcal{N} \setminus \{i, j\}}^{\bar{t} \setminus t}) \quad [8]$$

Residual information from variable i in the N -variable system \bar{X}_N^T is defined as follows.

$$R_{i,j}^T \equiv \sum_{t=1}^T I(X_i^t; X_j^t | \bar{X}_{\mathcal{N} \setminus \{i, j\}}^{\bar{t} \setminus t}) \quad [9]$$

Obviously, each of entropies and informations are non-negative: $H_i \geq 0$, $F_i \geq 0$, $T_{j \rightarrow i | \mathcal{N} \setminus \{i, j\}} \geq 0$, and $R_{i,j} \geq 0$ for arbitrary $i \in \mathcal{N}$ and $j \in \mathcal{N} \setminus i$. In a special case with a bivariate system ($N = 2$), it agrees with the bidirectional information network [7].

Properties of multivariate information network. The multivariate information network has the two major properties – holistic and local – stated in the two theorems. The first theorem states that, as a whole, a multivariate information network can be viewed as decomposition of total correlation among N variables [21, 22, 23]. The second theorem states that, at any node in the network, it holds Kirchhoff's current law or equivalence of the sum of in-coming information flow to the sum of out-going information flows (Figure 1d). In addition, each of information flows in the network is always non-negative, and this allows us to interpret each of informational quantities, entropy rate, free entropy, transfer entropy, and residual entropy, as an amount of information flow. The non-negativity of the flows is not trivial under simultaneous satisfaction of the Theorem 1 and 2 below. The present paper proves that MTE defined in Equations (6-9) has the theoretical properties stated in the following theorems⁴ (see also Supporting Information for the details).

Theorem 1 (Decomposition of total correlation). *In an N -variable system, total correlation consists of the sum of all the multivariate transfer entropies and residual entropies.*

$$C(\bar{X}_N^T) = \sum_{i=1}^N \sum_{j=1}^{i-1} G_{ij} \quad [10]$$

where $C(\bar{X}_N^T) = \sum_{i=1}^N H(X_i^T) - H(\bar{X}_N^T)$ is total correlation, and $G_{ij} = T_{i \rightarrow j | \mathcal{N} \setminus \{i, j\}} + T_{j \rightarrow i | \mathcal{N} \setminus \{i, j\}} + R_{i,j}^T$ is the sum of transfer entropies and residual entropies.

Theorem 2 (Local information flow). *Entropy rate of variable i consists of free entropy of variable i and the sum of in-coming multivariate transfer entropies to variable i as follows.*

$$H_i = F_i + \sum_{j \in \mathcal{N} \setminus i} T_{j \rightarrow i | \mathcal{N} \setminus \{i, j\}} \quad [11]$$

Entropy rate of variable i can be locally decomposed with the sum of all the in-coming and out-going multivariate transfer entropies and residual entropies.

$$H_i = H(\bar{X}_i^T | \bar{X}_{\mathcal{N} \setminus i}^T) + \sum_{j \in \mathcal{N} \setminus i} G_{ij} \quad [12]$$

³ This relationship between transfer entropy and mutual information has been pointed out originally in [7] without the residual term $R_{X,Y}$. In [6], the transfer entropy $T_{X \rightarrow Y}$ was defined as $T_{X \rightarrow Y} + R_{X,Y}$ in the current notations.

⁴ However, note that these theorems may not hold for an empirical MTE estimated with approximation (e.g., supposed Markov chain and/or stationary of time series) when a dataset violates the assumed approximations.

Numerical and empirical studies

As numerical and empirical validation of the MTE, we report simulation studies with the two classes of nonlinear dynamical systems and two case studies of empirical data analyses. Nonlinear dynamical systems would be one of interesting testbeds to demonstrate characterization of directed dependency structure of an unknown system with MTE. In a narrow sense, a generating process of a dynamical system is deterministic. Despite its determinism, its long-term chaotic behavior may be unpredictable, and it can be treated as a pseudo random series with knowledge of the initial state at finite precision. Yet it holds local dependency among variables at each time step. In each of the simulations, we generated a sufficiently long time series from a nonlinear dynamical system with a specific set of parameters. Then we tested whether we can recover the intrinsic dependent relationship on the basis of MTE or PTE applied to the generated time series.

In the empirical studies, we applied the MTE to two empirical datasets. One is a physiological dataset which has been analyzed in multiple previous studies in the context of a nonlinear time series analysis. The other is a dataset of human body movements in complex actions. Common in these two multivariate time series, complex systems in general need to coordinate multiple subcomponents in order to hold some intermediate states neither perfectly static, periodic, nor chaotic. Thus it is of great interest to analyze its mutual relationship among multiple subcomponents.

Simulation 1: Lorenz system. In Simulation 1, we applied the MTE to the Lorenz attractor which is one of well-known three dimensional dynamical systems defined with the following set of ordinary differential equations [24].

$$\begin{cases} \frac{dx}{dt} = \sigma(y - x) \\ \frac{dy}{dt} = x(\rho - z) - y \\ \frac{dz}{dt} = xy - \beta z \end{cases} \quad [13]$$

where x , y , and z is the system state, t is time, σ , ρ , and β are the system parameters. Given the ordinary differential equation, we suppose that each differential equation reflects information flows from the past states (x , y , and z) at t to the next states with a short time lag ($t + \Delta t$). Then the question here is whether we can infer these dependencies by applying information theoretic measures to the generated time series without prior knowledge of the differential equations. In the Lorenz equations, the differential $\frac{dy}{dt}$ and $\frac{dz}{dt}$ depends on all three variables, meanwhile the differential $\frac{dx}{dt}$ depends on x and y but not z . This asymmetric relationship – the variable z depends on x but not vice versa – gives a challenge to the measure of dependency in the multivariate system. Based on a good measure of dependency, we can reject the (conditional) dependence from x to z and detect the others.

Due to the Lorenz system being defined over continuous time, its time series was analyzed by manipulating the time lag Δt of the first order Markov chain $\{t, t + \Delta t\}$ systematically from 0.001 to 0.15. The upper panel and bottom panel in Figure 2 show the averages of MTEs and PTEs respectively as a function of the lag Δt . We performed statistical inference by taking the zero MTE or PTE (conditional independence) as the null hypothesis (See also Method for details), and the upper confidential bounds of the theoretical zero MTE and PTE are shown as a solid line.

The results showed that the MTE from variable z to x (highlighted in red), which is to be zero in theory, was evaluated as the lowest among the six directed pairs at all the lags except for $\Delta t < 0.01$ (Figure 2a). At the lags $0.06 < \Delta t < 0.07$, the to-be-zero MTEs from z to x were around the upper confidential bound of the theoretical zero MTE. Meanwhile, the MTE between the other five directed pairs were significant positive at any lags. The results suggested that

we could estimate the latent dependent relationship in the Lorenz system on basis of the MTE except for too short lag.

On the other hand, the results showed that the PTE from variable z to x was evaluated as the middle among the six directed pairs of them (Figure 2b), and it was significantly larger than the theoretical zero PTE at all the lags $0.001 < \Delta t < 0.15$. On the other hand, the PTE from y to x and from x to y , which should be positive in theory, tended to be as low as the theoretical zero PTE at all the lags. These results suggest that PTE does not just overestimate to-be-zero information flow but also underestimate to-be-positive information flows. In sum, these simulations suggest that the MTE may measure multivariate dependent structure with more accuracy than the PTE. The simulation clearly demonstrated the potential limitation of the PTE when applied to a system with the three variables or more.

Simulation 2: Characterization of various dependent structures.

Simulation 1 suggests the advantage of MTE and potential limitation of PTE in analysis of multivariate systems. In Simulation 2, we analyzed the robustness of MTE-based inference as a function of the number of variables, various topological types of dependent networks, and effects of unobserved noisy variables. Specifically we studied a class of *coupled map lattices* (CML) which allows us to systematically manipulate its parameters. The CML is a class of nonlinear dynamical system which linearly combines multiple one-dimensional chaotic systems as subcomponents [25]. Although each subsystem behavior is relatively simple and well known, it also shows a global emergent pattern across subsystems with a particular network topology among subsystems. Due to these properties, some variants of coupled map lattice have been used for modeling various kinds of real-world phenomena such as earthquakes [26], form of neurons [27], traffic flow [28], open flow [29], convection [30], cell-gene interaction [31], epileptic seizures [32] and so on. Utilizing its controllability, we analyzed the robustness of MTE based inferences on the system dependency.

Specifically, we studied a coupled tent map lattice (CTML) which is defined as follows. For $0 < x_i^t < 1$ ($i = 1, 2, \dots, N$ and $t = 0, 1, \dots, T$),

$$x_i^{t+1} = f \left(\frac{x_i^t + \epsilon \sum_{j \neq i} \delta_{ij} x_j^t + \eta_i^t}{1 + \epsilon \sum_{j \neq i} \delta_{ij} + \eta_i^t} \right) \quad [14]$$

where $\epsilon \geq 0$ is the coupling parameter indicating the degree of dependency in the system, δ_{ij} is either one or zero indicating existence of dependency from j to i , η_i^t is noise, a random value drawn from an uniform distribution, and $f(x)$ is the so-called tent map in which $f(x) = 2x$ if $x < \frac{1}{2}$, otherwise $f(x) = 2 - 2x$. As in Simulation 1, we define a positive coupling parameter between the variable x_i^t and x_j^{t+1} as positive information flow from variable i to variable j . Given binary information flows, either positive or zero, there are 16 different types of dependent networks with directed edges by identifying symmetric topology under exchange of the three variables (Figure 3a). Each of the 16 diagrams corresponds to a 3×3 matrix of network topology δ_{ij} ($i = 1, 2, 3, j = 1, 2, 3$) in Figure 3b. The colored (i, j)-cells in the matrices indicate the coupling parameters from x_j^t to x_i^{t+1} : white = 1, gray = ϵ , and black = 0 in Equation (14). Given the CTML, we systematically explored all the possible dependency diagrams with three, four, and five variables. There are 2^6 , 2^{12} , and 2^{20} possible combinations of inferences on binary information flows for each dataset, 96, 2616, and 192160 directed pairs in 16, 218, and 9608 unique diagrams for three, four and five variables respectively (Table 1).

Given the latent positive or zero information flows of the CTML, we generated the time series and computed MTE (PTE) for each directed pair. Then we defined MTEs (PTEs) larger than its 99%-confident upper bound of the theoretical zero MTEs (PTEs) as significant positive information flows. We analyzed the correspondence between the estimated and latent information flows as correct infer-

ence. Figures 3c and 3d show the results of estimated dependent pairs in the CTML with the three variables and the coupling parameter $\epsilon = 0.2$ without noise ($\eta_i^t = 0$ for any t). The estimated significantly positive MTEs or PTEs are shown as gray, otherwise black (See also Method for the statistical inference). The results showed that the MTE successfully gave correct inferences of dependency for all the 96 directed pairs in the 16 diagrams (Figure 3d). Meanwhile, the PTE overestimated the six directed pairs in four diagrams, and caused incorrect inferences which are highlighted in red in Figure 3c.

All the results of Simulation 2 including the CTML four and five variables are summarized in Table 1. The case-based correctness was defined as correct inferences for all the directed pairs in each combination of the diagrams. For the four-variable CTML, the proportion of correct inference based on the MTE was 90.78% of the 218 cases and 97.78% of the 2616 directed pairs. Meanwhile, that based on the PTE was 9.22% of the cases and 79.67% of the directed pairs. For the five-variable CTML, the proportion of correct inference based on the MTE was 81.48% of the 9602 cases and 95.41% of the 192160 directed pairs. That based on the PTE was 1.24% of the cases and 71.23% of the directed pairs. In sum, the simulation with the CTMLs demonstrated that robustness of MTE across the varying number of variables. On the other hand, inference based on the PTE tended to be more inaccurate as a function of the number of variables.

Robustness to unobserved variables

In the next analysis, we tested the robustness of the MTEs when applied to a dataset with unobserved variables. One of lessons derived from Simulation 1 is that PTE for a system with three or more variables may be inaccurate. The same lesson could apply to N -variable MTE for ones with $(N + 1)$ or more potential variables. Since, unlike the simulations, we cannot always observe all the sufficient set of variables in typical empirical data analyses, it raises a potential concern that MTE may not be better than PTE for the dataset with potential unobserved set of variables. Therefore, it is of practical importance to evaluate robustness of the measures for such datasets with unobserved variables. In the simulation, we generated the time series following the equation X with the noise term η_i^t is a random value drawn from the uniform distribution $[0, 0.1]$ for each time step t . The random inputs η_i^t to the variable x_i^{t+1} reflects perturbation by the unobserved set of variables. With or without noisy unobserved variables, we analyzed the average classification performance of the MTE and PTE for the 16 cases of 3-variable CTMLs (Figure 3a) as a function of the coupling rate ϵ from 0 to 0.6 (Figure 3a and 3b).

In analysis of the dataset without unobserved variables, on the basis of TE, we made correct inferences in all the cases at a small range of coupling parameters $0.05 \leq \epsilon \leq 0.1$ (Figure 4a). Likewise, on basis of MTEs, we made correct inference in all the cases at a relatively broader range of coupling parameters $0.1 \leq \epsilon \leq 0.25$ (Figure 4b). Since too large coupling parameters ($\epsilon > 0.25$) made multiple variables perfectly coupled ($R > 0.95$), those coupling variables were difficult to discriminate on the basis of MTEs with its coarse-grain encoding of the phase space. In fact, we found at least one false detection due to nearly perfect coupling in Cases 10, 11, 13, 14 and 15 (Figure 3a). In contrast, the advantage of PTEs in a small coupling parameter is likely to be caused by the relatively small effects of the third variable. In addition, MTE needs to estimate the probabilistic distribution a large combinatorial space relative to the given sample size. This sparseness of samples leads MTE to be more conservative to reject false positive information flows. Except for this small coupling rate advantage, MTE outperformed PTE in most of the cases and parameters.

Simulation of the datasets with noisy latent variable showed basically similar patterns as found in that of the noiseless dataset. In Figure 4b, we found the advantage of PTE in a small coupling parameter $\epsilon = 0.1$ or smaller and the better detection performance of MTE otherwise. We also found a remarkable difference from the noise-

less dataset: MTE tended to show even better performance while TE tended to show worse performance in the dataset with noisy latent variable. In this particular simulation, the reason for the even better performance of MTE for the noisy data was perhaps because the noisy variables decoupled the perfect-coupled variables (at $\epsilon > 0.25$) which harmed MTE performance for the noiseless dataset. In sum, these results suggested the robust relative advantage of MTE even with noisy latent variables.

Summary of numerical studies

We summarize the findings in Simulations 1 and 2 in the four points. First, MTE showed advantages over PTE in both continuous-time system asymmetric under exchange of variables (Simulation 1) and discrete-time systems (Simulation 2). Second MTE also showed advantages in various types of dependent topology in the CTMLs with 3, 4, and 5 variables. Third, its advantage is robust even in the analysis of the datasets with unobserved noisy variables. Finally, the simulations also showed a limitation of the MTE, more conservative estimate of information flows than PTE. We will discuss this technical limitation in the later section.

Empirical data analysis 1: Physiological data. As a case study, we analyzed a physiological dataset including three vital signs recorded in a sleeping person [33, 34]. Besides being a trivariate time series, we chose this dataset as a benchmark test, since it has been analyzed across many theoretical studies [2, 3, 35, 36]. The original data consists of the set of three time series of heart rates, breath rates, and blood oxygen concentration recorded at 0.5 Hz of sampling rate. The particular person measured has been known to show respiratory sinus arrhythmia. It is a frequently-seen symptom that shows correlation between heart rates and breath rates. As expected, the previous study showed that the heart rates and breath rates transferred information bidirectionally by applying PTE [3]. However, as suggested in Simulation 1 and 2, it is potentially possible to have such seeming information transfers caused by the third factor, for example, the blood oxygen concentration in this dataset. Thus, we performed re-analysis on the dataset not just as bivariate but as a part of a trivariate system by applying the MTE.

Figures 5a and 5b respectively show the PTE (as bivariate series) and MTE (given the blood oxygen concentration) between heart rates and breath rates as a function of time lag. In Figure 5b, we replicated the qualitative patterns of PTEs as found in the previous study: both directions have information transfers at most of time lags, while the heart rate tended to transfer information to the breath rates more than the other direction⁵.

The qualitative patterns of PTE and MTE basically agreed - heart rates and breath rates are tightly coupled bidirectionally with or without respect to blood oxygen concentration. This result confirmed the conclusion in the previous study even as a part of a trivariate system in regard to these qualitative patterns. However, we also found a difference between the two measures. In MTE, we also found that information transfer between the heart rates and the breath rates peaked around the same time scale of the lag approximately 2 sec. One the other hand, in PTE, the two directions had peaks at quite different scales of time lags: PTE from Heart to Breath peaked at approximately 2 sec and that from Breath to Heart was at approximately 20 sec. At this moment, we could not conclude which of the results, synchronized or delayed peaks in MTE and PTE, is more plausible in light of empirical findings. It is an open question for further empirical studies.

Empirical data analysis 2: Motor coordination in complex actions. The second case study is an analysis of complex human actions. Our

⁵ It is potentially possible to have the results disagreed in the present and previous study due to its technical difference in estimation method and choice of a particular subset of time series.

bodily actions require coordinated movements of multiple body parts. A human body consists of over two hundreds bones, numerous muscles, and billions of neurons in central and peripheral nerve systems controlling them with feedback loops. Obviously, making a smooth action requires integrated control over all levels of these systems. It is of our interest to characterize human motor coordination in skillful actions through the MTE. Specifically, we chose a dataset of complex actions performed by multiple players with different expertise levels. The data was originally obtained in order to analyze the levels of expertise in the samba music plays [37, 38]. The original dataset consists of five players, and each player performed basic samba shaking actions in five different tempos (60, 75, 90, 105, 120 beats per minute, and each trial lasted 97.4 seconds on average) by being cued with a metronome. While playing, three dimensional motions of 18 markers, attached on body parts and musical instruments, were recorded at 86.1Hz of sampling rate.

As well as the original study, here we aim to find the relationship between informational properties among bodily actions and the expertise levels in the motor skill. The present study analyzed the actions of three players, chosen from the original five, who are one master player (more than thirty years of experience) and two of his disciples Disciple 1 (six years of experience) and Disciple 2 (two years of experience). The expertise levels between the master and his disciples were expected to be different. Given our knowledge of the players' expertise levels as the ground truth, we tested whether MTE can successfully detect the differences in their skill levels. For simplicity, we limited ourselves to analyze a subset of the original datasets, 3190 samples (74.1 seconds long) of four motions of markers attached on right wrist, right elbow, and two sides of the musical instrument (shaker). These were the essential parts of the samba actions making sounds directly, and we expected that information flows among them would be crucial to characterize the players' expertise levels. In a smooth samba play, multiple body parts need to be coordinated to perform the complex actions. Thus, perhaps common in general multivariate time series analyses, one of challenges in this analysis is to decompose the smooth actions into information flows between body parts.

In the analysis, we applied the MTE and PTE to all the directed pairs of four motions. Figure 6 shows the proportion of directed pairs with significantly positive MTE and PTE averaged across five different tempo conditions ($60 = 12 \times 5$ directed pairs in total for each subject). The results showed distinguishable patterns of informational coupling among the master and the two disciples. Across all the five tempo conditions, we found all the body parts in the master player nearly perfectly coupled. Meanwhile, Disciple 2 with the least experience among the three showed the least number of informational coupled pairs. In Figure 6, each graph on the top shows the information network of each player. It has a solid edge between the markers if at least one of the two directions had significant positive MTE across all the five tempo conditions. The graph of Disciple 2 shows that the only consistently coupling pair was his wrist and a side of shaker. That of Disciple 1 showed the three edges, elbow-wrist, wrist-shaker2, and shaker2-shaker1, which suggests these physically connected parts formed a action like whip stroking. As expected, these results showed consistency between the player's expertise levels and the MTE-based informational properties in their bodily actions.

As a comparison, we also applied PTE to the same dataset (Figure 6). The results showed that PTE did not detect the differences between the master and Disciple 1 both of whom showed significant PTEs in all the directed pairs. Compared with the MTE, PTE tended to overestimate the coupling pairs in all three players. In the graph patterns, PTE detected positive information between the pairs which MTE did not detect. Regarding our knowledge of subjects' expertise levels, PTE estimation was likely to detect false positive information flows due to the effects of the two other unconsidered variables. As a

result, we could not find the difference in the PTEs among players as clearly as found in the MTEs. These results of empirical data analyses suggested potential applicability of MTE to empirical complex multivariate time series.

Discussion

The present study proposes an extended information theoretic measure for system characterization through time series. The multivariate transfer entropy is a natural generalization of pairwise transfer entropy to a multivariate system holding a set of theoretical properties. The MTE was tested on the two classes of nonlinear dynamical systems and on the two empirical datasets. The simulations demonstrated the advantage of MTE over PTE, in both discrete-time and continuous-time systems, with most of the topologies of dependency among 3, 4, and 5 variables, and even with additional noisy latent variable. These advantages of MTE would stem from the theoretical property that the MTE decomposes higher order dependencies into information flows in a network. Since the PTE does not always satisfy it for a system with three or more variables, the PTE from A to B may take some value independently of the PTE from A to C . As a result, the PTE from A to B may overestimate or underestimate dependency between A and B when the third variable C also has effects on B .

Application to the two empirical datasets suggested its potential use in the analysis of empirical complex systems including physiological signals and human motor coordination. In analysis of such datasets with complex interactions, MTE is likely to be useful because it exclusively measures a pair of variables by cancelling out the effects from the other variables. This general applicability of MTE to multivariate systems covers a broader range of empirical and theoretical fields using PTEs [8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20].

Technical limitations and future works. In contrast to its relatively robust and accurate evaluations, the present simulation studies also suggested a limitation of the MTE. The analysis in Simulation 2 showed that the MTE was more conservative to detect dependency than the PTE. This problem of MTE was likely to be caused by the technical issue in its estimation. The N -variable MTE needs conditional entropy of N variables in which the combinatorial space may grow as an exponential function of the number of variables. It causes high computational costs and inaccuracy of the estimation due to the sparsity of samples relative to the exponentially growing space. In the current implementation, which was not optimized but was computed in a naive way, it was very costly to compute even relatively modest number of variables $N > 5$. This estimation problem prevented us from using a finer-grained binning on the phase space, and we suggest that it resulted in the conservative detection of information by MTE. Therefore, further work could include developing a technique relaxing this problem. Similar technical issues have been discussed for PTE such as small-sample correlation for pairwise transfer entropy [15] or non-parametric probabilistic density estimator for continuous time series [39].

Another related concern in empirical analyses is parameter specification. In order to accurately measure dependency, we need to specify temporal delay and estimators of probabilistic distributions. The current simulations were demonstrated with one of the simplest probabilistic models - binary coding by median splitting with the first order Markov chain. It is an open question to what extent estimation of the MTE depends on the choice of these parameters which perhaps depends on the case. More importantly, how can we choose the probabilistic model and delays? One of potential solutions for this problem for dynamical systems is *generating partitions*. A set of generating partitions gives a theoretical ground for "best" discrete states of discrete or continuous dynamical system which has one-to-one correspondence between a series of symbols and a subset of state space. It has been constructed for several low dimensional chaotic

systems (e.g., coupled map lattice [40] and Henon map [41]) and several algorithms estimating symbolic dynamics from an empirical time series have been proposed [42, 43, 44].

Methods

In all the analyses in the present study, a set of continuous time series of T samples of N variables was converted to a symbolic form encoding the original data in a coarse grained representation (See each analysis below for the specific symbolization process). Based on the symbolic series, probabilistic distribution of time series assuming K -th order Markov chain was estimated, then MTE and PTE were estimated on the probabilistic distribution subsequently (See Estimation of probabilistic distribution below for details). The estimated MTE (PTE) was compared to the corresponding theoretical zero MTE (PTE) (See Zero transfer entropy below). All the computational routines used in the simulations and empirical data analyses were written on the MATLAB platform, which is available at the author's website (<http://www.jaist.ac.jp/~shhidaka/>).

Estimation of probabilistic distribution. In each simulation above, an array of the values $x_{i,t}$ for $i = 1, 2, \dots, N$, $t = 1, 2, \dots, L$ was given as the N -dimensional time series of length L . Each value in dimension i was symbolized $\hat{x}_{it} = f(x_{it})$ where $f_M(x) : R^1 \rightarrow \mathcal{M}_i^1$ is the symbolization function mapping the one dimensional real space to the symbol set of M_i alphabets, $\mathcal{M}_i = \{1, 2, \dots, M_i\}$, which is specified in each simulation. The N -tuple, $y_t = \{\hat{x}_{1,t}, \hat{x}_{2,t}, \dots, \hat{x}_{N,t}\} \in \mathcal{M}^N$, was treated as the joint space of N symbols. Assuming stationarity and K -th order Markovity, the conditional probabilistic distribution $P(y^t | y^{t-1}, y^{t-2}, \dots, y^{t-K})$ over $\{t - K, t - K + 1, \dots, t\}$ was estimated by its maximum likelihood estimator (MLE) assuming the multinomial distribution over the joint symbol space $\{\mathcal{M}_1^K \otimes \mathcal{M}_2^K \otimes \dots \otimes \mathcal{M}_N^K\}$. Specifically, the MLE is the frequency over the joint symbol space $\{\mathcal{M}_1^K \otimes \mathcal{M}_2^K \otimes \dots \otimes \mathcal{M}_N^K\}$ which is normalized to be probability. According to the stationary assumption, $P(y^{t_1} | y^{t_1-1}, y^{t_1-2}, \dots, y^{t_1-K}) = P(y^{t_2} | y^{t_2-1}, y^{t_2-2}, \dots, y^{t_2-K})$ for any $t_1 \leq L - K$ and $t_2 \leq L - K$. Thus, a series within the time window of $K + 1$, $z_t^{t+K} = \{y_t, y_{t+1}, \dots, y_{t+K}\}$, was counted across the data length $t = 1, 2, \dots, L - K$, thus, a dataset of length L provides $L - K$ samples for estimation of the conditional probabilistic distribution $P(y^t | y^{t-1}, y^{t-2}, \dots, y^{t-K})$. Since the joint symbol space has a large number of possible combinations $\prod_i^N M_i^K$ growing as exponential function of dimension N and time window length K , an empirical dataset with a limited sample size may be too sparse to estimate probabilistic distribution over the joint symbolic space. Therefore, another reasonable choice for a sparse-data estimator would be the ones combined with various kinds of smoothing techniques on the K -gram model. The author confirmed that the modified Kneser-Ney smoothing [45] was effective in particular for data with small sample size, although the present paper reported the MLE estimator in all the analyses.

Zero transfer entropy. The estimate transfer entropy was compared with the null hypothesis that the true (pairwise or multivariate) transfer entropy is zero meaning conditional independence of a given pair of variables. Note that, even under such the null hypothesis, estimated transfer entropy may be positive due to a finite sample size for estimation. The probabilistic distribution of the null transfer entropy (as a special case of conditional mutual information) follows gamma distribution with the shape parameter s_1 and the scale parameter s_2 [46]. The parameters s_1 and s_2 may vary across simulations due to specific features of samples, but its maximum is X and Y in Simulation 1 and Z and W in Simulation 2, XX and YY in the empirical data analysis.

Simulation 1. In Simulation 1, we generated a time series from the Lorenz system from $t_0 = 50$ to $t = 2050$ with the initial value $\{x(t_0), y(t_0), z(t_0)\} = \{1 + \eta_1, 1/2 + \eta_2, 0 + \eta_3\}$ and the parameters $\{\sigma, \rho, \beta\} = \{10, 28, 8/3\}$ (Equation 13) where each of the noise factors $\{\eta_1, \eta_2, \eta_3\}$ is a random value drawn from uniform distribution from 0 to 0.01. Using the solver of the ordinary differential equation (ode45 routine in the MATLAB), we obtain approximately 118,000 samples of the three dimensional series, and resample the time series by linear interpolation as desired temporal resolution from 0.001 to 0.15 per sample. For each of given temporal resolutions Δt , we obtained 300,000 samples by taking a subset of c sets of the time series with different initial values where $c = \lceil \frac{300000\Delta t}{2000} \rceil$ and $\lceil x \rceil$ is the maximum integer equal or smaller than x . Given a set of 3 dimensional time series, in order to form probabilistic distribution, we convert each variable to binary series s_i^t of dimension i and time t by splitting median point of each dimension. The first order Markov chain $p(s_i, s_i^{t+\Delta t})$ (for $i = 1, 2, \text{and } 3$) was used for MTE and TE estimation.

Simulation 2. We generate time series x_i^t of the coupled tent map lattice based on the Equation 14 with a given set of parameters (coupling parameter ϵ and δ_{ij} for $i = 1, 2, \dots, N$, $j = 1, 2, \dots, N$ in Equation 14 of N variables), a set of random values η_i^t drawn from $(0, r)$ r is either 0 or 0.1, and a set of random initial values x_i^1 drawn from uniform distribution of the range $(0, 1)$. In each case, a time series of 10^5 steps after the first 1000 samples discarded as transient was converted to binary values s_i^t by median splitting $s_i^t = h(x_i^t \geq \bar{x}_i^t)$ where \bar{x}_i^t is the median of x_i^t ($i = 1, 2, \dots, 10^5$) and the Heaviside function $h(x)$ is 1 if x is positive 0 otherwise. The third order Markov chain $p(s_i, s_i^{t+2}, s_i^{t+3}, s_i^{t+4})$ (for $i = 1, 2, 3$) was used for MTE and TE estimation.

Empirical data analysis 1. We analyzed the dataset B of the trivariate time series, heart rates, breath rates, and blood oxygen concentration, retrieved from the Santa Fe time series competition (<http://www-psychem.stanford.edu/~andreas/Time-Series/SantaFe.html>). We concatenated all the consecutive time series longer than 250 seconds and in the waking states diagnosed by the expert, and made a trivariate time series of 11560 samples. For each of given temporal resolutions Δt , we obtained 20,000 samples by taking subset of C sets of the time series with different lags $\{t_0 + t, t_0 + t + \Delta t, \dots, t_0 + t + k\Delta t\}$ where $t = 0, \frac{\Delta t}{c}, \dots, \frac{(c-1)\Delta t}{c}$, $c = \lceil \frac{20000\Delta t}{11560} \rceil$ and $\lceil x \rceil$ is the maximum integer equal or smaller than x .

Empirical data analysis 2. The dataset consists of three players performing in five different tempos (60, 75, 90, 105, 120 beats per minute, and each trial lasted 97.4 seconds on average). The movements of four markers attached on right elbow, right wrist, and the two sides of musical instruments, each originally recorded at 86.1 Hz, were analyzed. In the analysis, after down-sampling the original data to 46.05 Hz, the first 250 samples (5.81 second long from the beginning of the recording) were excluded as initial setup of the actions, and 3250 samples (75.5 second long) of movements were analyzed for each subject. In order to reduce measurement noise, for each movement of the markers, the local linear projective method was performed after phase space reconstruction of each time series on the 31 dimensional time delay space with 46 msec (i.e., $\{t, t + \Delta t, t + 2\Delta t, \dots, t + 30\Delta t\}$ where $\Delta t = 46$ msec) [2]. For each estimated phase space, a symbol series was assigned using the symbolic false nearest neighbor method [43] which estimates a generating partition for a time series. Given the four-variable symbolic series, we applied the multivariate transfer entropy. The estimated MTE greater than the zero MTE at the level of $p < 0.001$ were defined as a significant MTE for each of the four body parts in each condition.

ACKNOWLEDGMENTS. The author is grateful to Dr. Tsutomu Fujinami for his kind offering of his dataset of a set of complex human actions. The author thanks Dr. Neeraj Kashyap, Dr. Brian Kurkoski, Takuma Torii, and Akira Masumi for their

helpful discussions and comments on the early version of the manuscript. This work was supported by Artificial Intelligence Research Promotion Foundation and Grant-in-Aid for Scientific Research B No. 23300099.

1. Gershenfeld, N. A. & Weigend, A. S. (1993) *The Future of Time Series*, eds. Weigend, A. S. & Gershenfeld, N. A. (Westview Press, Boulder, CO).
2. Kantz, H. & Schreiber, T. (1997) *Nonlinear Time Series Analysis*. (Cambridge University Press, Cambridge, UK).
3. Schreiber, T. (2000) *Phys. Rev. Lett.* 85, 461–464.
4. Shannon, C. E. (1948) *The Bell System Technical Journal* 27, 379–423, 623–656.
5. Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory*. (Wiley-Interscience), 99th edition.
6. Massey, J. L. (1990) *Causality, Feedback and Directed Information*.
7. Marko, H. (1973) *IEEE Transaction on Communication COM-21*, 1345–1351.
8. Lizier, J. T., Prokopenko, M., & Zomaya, A. Y. (2008) *Phys. Rev. E* 77, 026110.
9. Quinn, C., Coleman, T., & Kiyavash, N. (2010) *Approximating discrete probability distributions with causal dependence trees*. pp. 100–105.
10. Bauer, M., Cox, J. W., Caveness, M. H., Downs, J. J., & Thornhill, N. F. (2007) *Control Systems Technology, IEEE Transactions on* 15, 12–21.
11. Nichols, J. M., Seaver, M., Trickey, S. T., Todd, M. D., Olson, C., & Overbey, L. (2005) *Phys. Rev. E* 72, 046217.
12. Rubinov, M. & Sporns, O. (2010) *NeuroImage* 52, 1059–1069. Computational Models of the Brain.
13. Vicente, R., Wibral, M., Lindner, M., & Pipa, G. (2011) *J. Comput. Neurosci.* 30, 45–67.
14. Baek, S. K., Jung, W.-S., Kwon, O., & Moon, H.-T. (2005) *ArXiv Physics e-prints*.
15. Marschinski, R. & Kantz, H. (2002) *European Physical Journal B* 30, 275–281.
16. Moniz, L., Cooch, E., Ellner, S., Nichols, J., & Nichols, J. (2007) *Ecological Modelling* 208, 145–158.
17. Kulp, C. W. & Schlingmann, D. (2009) in *Mathematics and Computation in Music, Communications in Computer and Information Science*, eds. Klouche, T. & Noll, T. (Springer Berlin Heidelberg) Vol. 37, pp. 441–448.
18. Hidaka, S. & Yu, C. (2010) *Spatio-Temporal Symbolization of Multidimensional Time Series*. pp. 249–256.
19. Hidaka, S. & Yu, C. (2010) *Analyzing multimodal time series as dynamical systems, ICMI-MLMI '10*. (ACM, New York, NY, USA), pp. 53:1–53:8.
20. Sumioka, H., Asada, M., & Yoshikawa, Y. (2007) *Causality detected by transfer entropy leads acquisition of joint attention*. pp. 264–269.
21. Watanabe, S. (1960) *IBM J. Res. Dev.* 4, 66–82.
22. Garner, W. R. (1962) *Uncertainty and Structure as Psychological Concepts*. (John-Wiley & Sons, New York: NY).
23. Studený, M. & Vejnarová, J. (1999) *The multiinformation function as a tool for measuring stochastic dependence*, ed. Jordan, M. I. (MIT Press, Cambridge, MA).
24. Lorenz, E. N. (1963) *Journal of Atmospheric Sciences* 20, 130–148.
25. Kaneko, K. (1992) *Chaos* 2, 279–282.
26. Ceva, H. (1995) *Phys. Rev. E* 52, 154–158.
27. Sakaguchi, H. & Ohtaki, M. (1999) *Physica A: Statistical Mechanics and its Applications* 272, 300–313.
28. Yukawa, S. & Kikuchi, M. (1995) *Journal of the Physical Society of Japan* 64, 35–38.
29. Willeboordse, F. H. & Kaneko, K. (1995) *Physica D: Nonlinear Phenomena* 86, 428–455.
30. Yanagita, T. & Kaneko, K. (1993) *Physics Letters A* 175, 415–420.
31. Bignone, F. A. (1993) *Journal of Theoretical Biology* 161, 231–249.
32. Larter, R., Speelman, B., & Worth, R. M. (1999) *Chaos* 9, 795–804.
33. Rigney, D., Goldberger, A., Ocasio, W., Ichimaru, Y., Moody, G., & Mark, R. (1993) *Multi-channel physiological data: description and analysis* eds. Weigend, A. & Gershenfeld, N. pp. 105–129.
34. Ichimaru, Y. & Moody, G. (1999) *Psychiatry and Clinical Neurosciences* 53, 175–177.
35. Angelini, L., Maestri, R., Marinazzo, D., Nitti, L., Pellicoro, M., Pinna, G. D., Stramaglia, S., & Tupputi, S. A. (2007) *Artificial Intelligence in Medicine* 41, 237–250.
36. Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008) *Phys. Rev. Lett.* 100, 144103.
37. Yamamoto, Y., Ishikawa, K., & Fujinami, T. (2006) *Journal of biomechanics* 39, S555.
38. Yamamoto, T. & Fujinami, T. (2008) *Human Movement Science* 27, 812–822.
39. Kaiser, A. & Schreiber, T. (2002) *Physica D: Nonlinear Phenomena* 166, 43–62.
40. Pethel, S. D., Corron, N. J., & Bollt, E. (2006) *Phys. Rev. Lett.* 96, 034105.
41. Biham, O. & Wenzel, W. (1990) *Phys. Rev. A* 42, 4639–4646.
42. Kennel, M. B. & Buhl, M. (2003) *Physical Review Letters* 91, 084102.
43. Buhl, M. & Kennel, M. B. (2005) *Physical Review E* 71, 046213.
44. Hirata, Y., Judd, K., & Kliminster, D. (2004) *Physical Review E* 70, 016215.
45. Chen, S. F. & Goodman, J. (1999) *Computer Speech & Language* 13, 359–393.
46. Goebel, B., Dawy, Z., Hagenauer, J., & Mueller, J. (2005) *An approximation to the distribution of finite sample size mutual information estimates*. Vol. 2, pp. 1102–1106 Vol. 2.

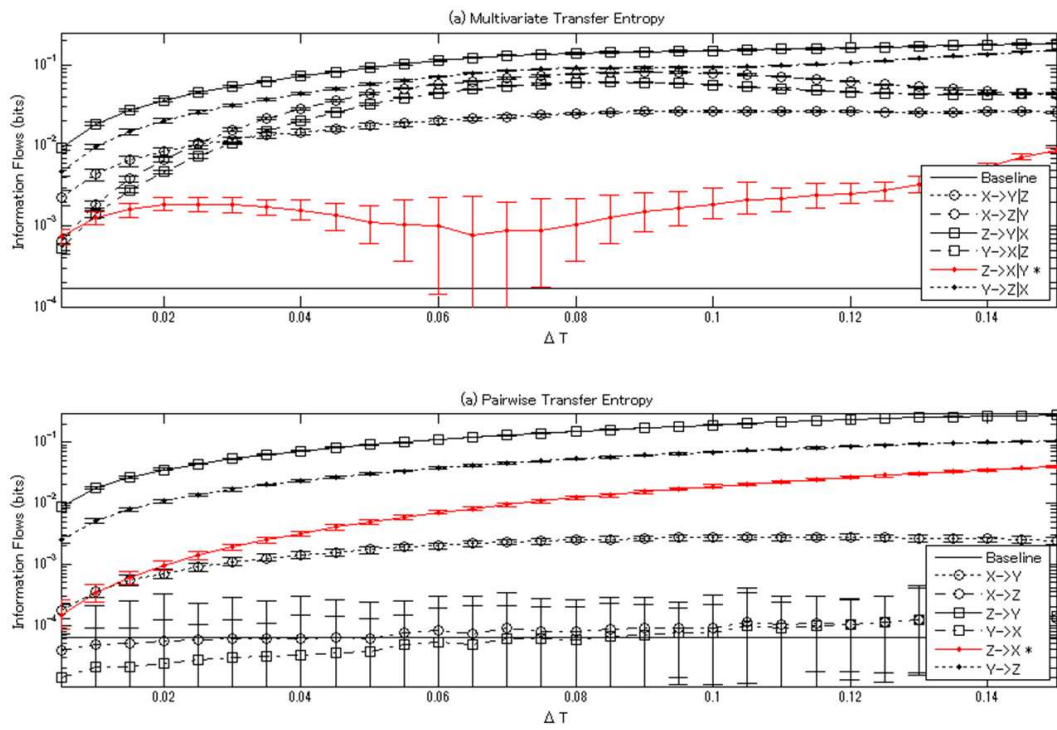


Fig. 2. The estimated PTE and MTE for Lorenz attractor as a function of time lag Δt . The directed pair from z to x to be zero in theory is highlighted in red.

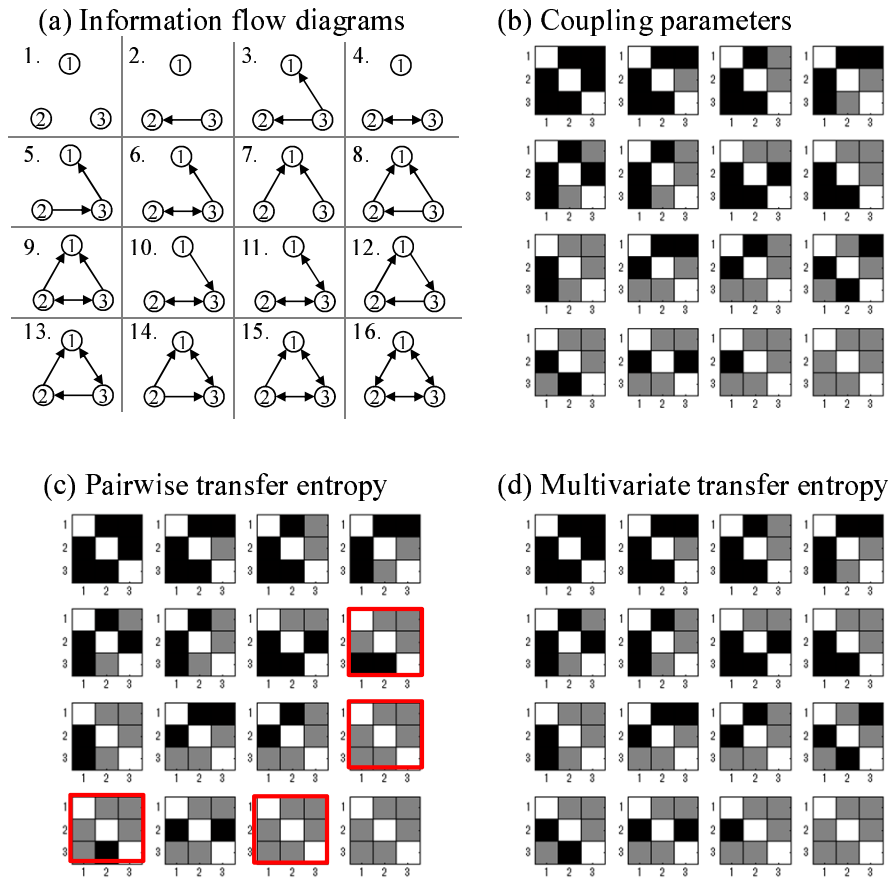
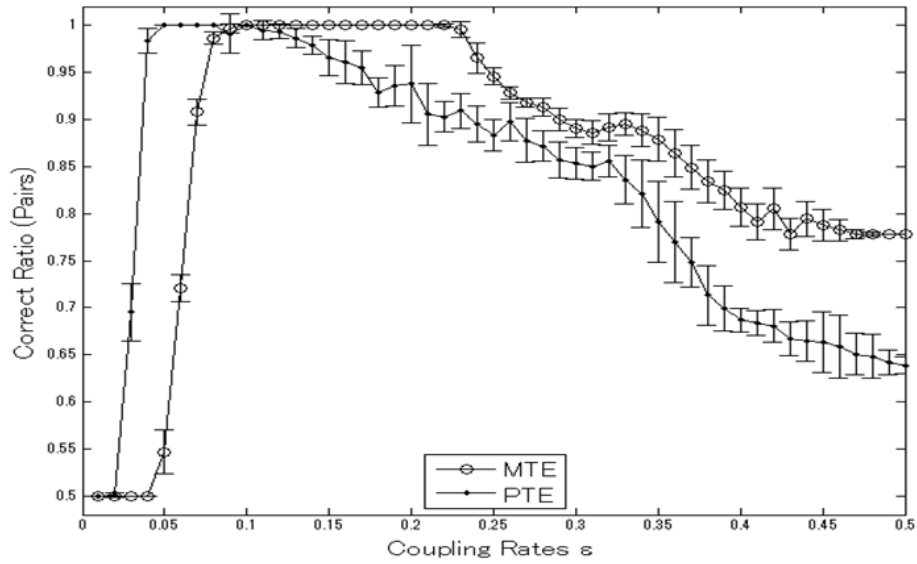


Fig. 3. (a) 16 possible diagrams of information flows (its direction indicated by the arrows) among the three variable 1, 2, and 3. (b) The matrices of coupling parameters in the coupled tent map lattice corresponding to the 16 information flow diagrams (white diagonal cell = 1, gray = ϵ , and black = 0), (c-d), The matrices of estimated information flows in which the significant positive PTE and MTE are in gray, the ones not significant are in black, the ones not tested (diagonal cells) are in white. The red outlines of matrices highlight cases with misdetection of information flows.

(a) CTML without noise (N=3)



(b) CTML with noise (N=3)

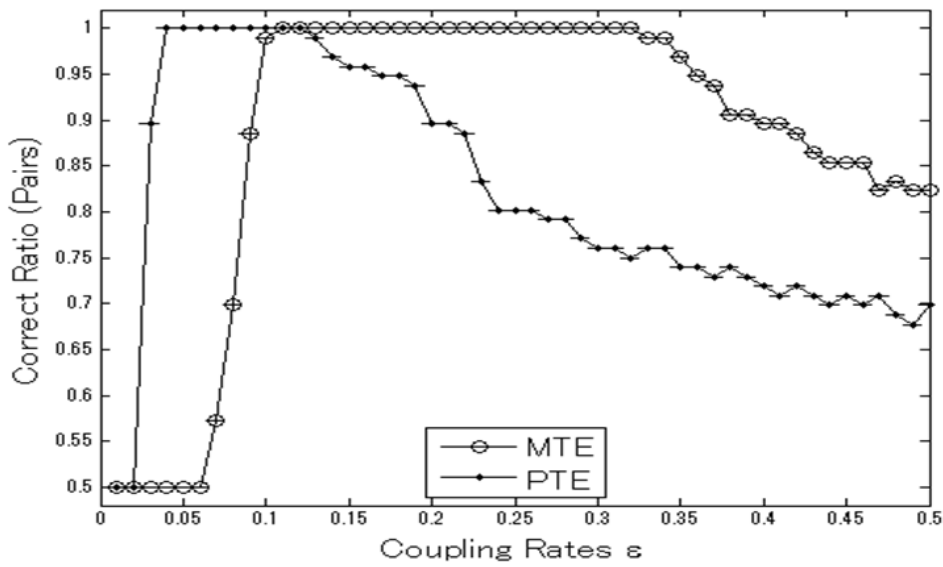


Fig. 4. The correct detection of information flows (directed pairs as its unit) as a function of the coupling parameter averaged across the 16 cases of coupled tent map lattice with three variables including (a) no noise factor ($\eta_i^t = 0$) and (b) noise factors ($0 \leq \eta_i^t \leq 0.1$).

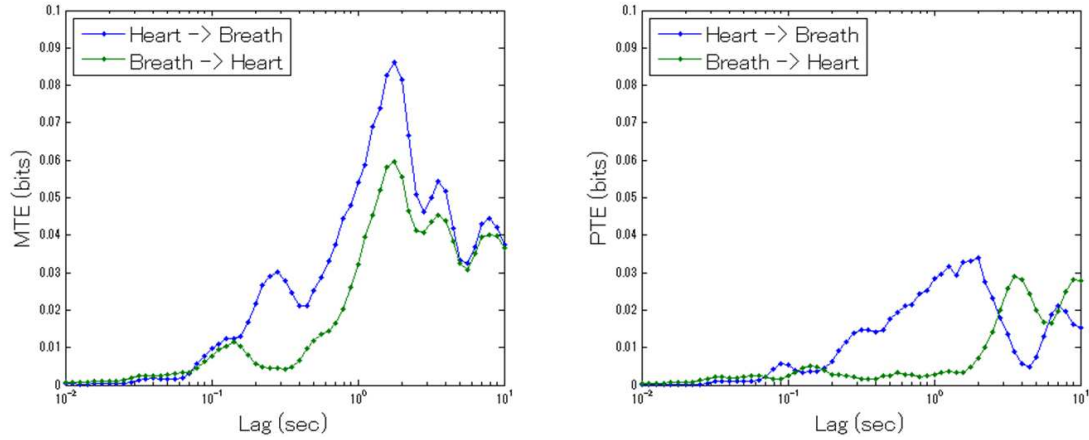


Fig. 5. The estimated information transfer between heart rates and breath rates with MTE given blood oxygen concentration (left) and PTE (right).

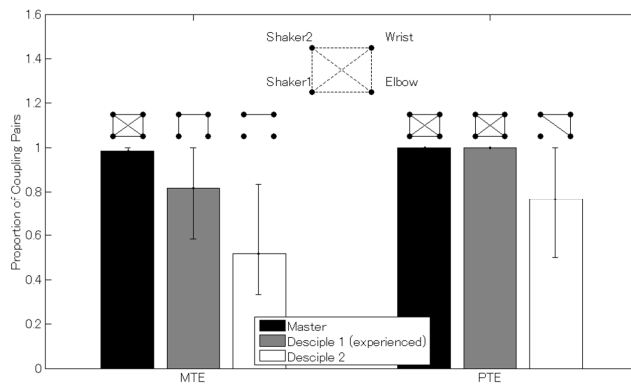


Fig. 6. the proportion of coupling bodily movements estimated with MTE and PTE. The proportions for each subject are average across 5 different tempo conditions, and its maximum and minimum are shown as the errorbar. Each graph on top of the bar shows informational connectivity among four bodily movements in which an edge shows one of either or both directions have significant MTEs or PTEs across all the five conditions.